Background

Breastfeeding is an important part of infancy, with breastmilk acting as both a nutritional source and as protection against potential infections during that period of time. Studies have shown that breast milk is also the strongest determinant of gut microbiota composition, and as a result can have lasting positive impacts on both the infant's metabolism and immune system [1]. The composition of human breastmilk contains important molecules called human milk oligosaccharides, which microbes such as *Bifidobacterium* rely on for survival. Although many other microbe-host interactions in humans still lack information, this symbiotic relationship based on HMOs was originally discovered in mice. It was established that *Bifidobacterium* presence in the guts of infants has significant benefits, therefore the study was interested in gaining insight on the specific mechanisms relating to metabolites allowing the bacterium to be beneficial.

The study was based on infants from the SKOT I cohort, with 30 males and 29 females recruited randomly from two regions in the National Danish Civil Registry. All of the children were around 9 months in age at the beginning of the study. Throughout the study, fecal samples from the infants were obtained and frozen for future DNA extraction.

The first portion of the study was targeted at learning how breastfeeding and the intake of breastmilk affected the metabolites present in infant microbiomes. In order to find the concentrations of the different aromatic amino acids and their derivatives in the fecal samples, an enzymatic assay was performed for each of the samples. Afterwards, HPLC analysis was used on the resulting reaction products in order to determine the concentrations.

After identifying the present metabolites and their concentrations, the study was interested to discovering if any of the metabolites produced from breastmilk impacted the concentrations of *Bifidobacterium* in the infant microbiomes. In order to obtain an estimate of the abundances of each of the microbes, quantitative PCR and 16S rRNA gene amplicon sequencing were used. qPCR using universal primers provided an estimate of the absolute abundances of the microbes, and 16S rRNA gene amplicon sequencing provided the relative abundances of microbes within the samples. When these two values are multiplied together, it results in an estimate of the absolute abundance of each of the microbial taxes. These absolute abundance estimates can then be used to discovery any correlations between concentrations and metabolite present.

All the data was stored in excel files in table 1 and table 2 of the supplementary materials.

Data Analysis and Structure

For figure 1c, the data was prepared by downloading table 1e as a csv file, and converting the contents into a nested list. The data contains fecal concentrations of aromatic amino acids and their derivatives in the SKOT cohort study. After converting all the values from strings to floats, I used pandas to create a data frame. I set sample numbers as the row labels, and the amino acid names as the column labels. In order to perform principal component analysis on the data, it needed to be normalized first. To do this, I first separated out the features and used StandardScaler from SciKit-learn to standardize them. After normalizing the data, I used SciKit-learn PCA to perform 2 component principal component analysis on the data. Since there were no targets used in the PCA, I added a column for breastfeeding status to the PCA results corresponding to each SKOT cohort study participant. This information was taken from table 1a, a metadata of the study participants containing information in column D whether or not infants were still undergoing partial breastfeeding. If they were still undergoing partial breastfeeding,

'Breastfed' was used, and if they were no longer partially breastfeeding, 'Weaned' was used in the new column. After initializing the targets, I used matplotlib pyplot to plot the principal components, labeling 'Breastfed' points in red and 'Weaned' points in blue.

In order to visualize which amino acids and their derivatives contribute the most to the two principal components, I created a variable factor map. I used pyplot to plot a circle, with arrows originating from the center for each variable. The coordinates of the endpoints of the arrows are taken from PC1 and PC2 values respectively. Therefore, the longer the arrow, the higher contribution the variable would have on the principal components. In order to simplify the plot for legibility, only the three highest contributors were identified and plotted.

For figure 1d, the same amino acid data from before was combined with information in table 1d on the different genus relative abundances in the SKOT cohort. In specific, the figure is looking at the abundance of *Bifidobacterium*, so I took that column of data and converted it into a list. In order to find the spearman's constant for each of the amino acids, I set the column labels as features, and for each feature used spearmanr from scipy.stats. This function takes the fecal concentrations for each feature and the relative abundances of *Bifidobacterium*, and calculates the rho and p-values for each feature. I saved both the rho and p values for all the features in two different lists, and combines them into a data structure after, changing the row names to the names of the features. In order to visualize the results, I used seaborn to create a heatmap corresponding to the spearman's rank values. In order to specify between the different types of aromatic amino acids and their derivatives, I also separated the heatmaps based upon their categories. There is a section for tyrosine metabolites, phenylalanine metabolites, and Tryptophan metabolites.

Conclusion

Figure 1c aimed to see if breastfeeding status had any impacts on the fecal amino acid metabolite concentrations in infants. Plotting the first two components of the PCA results show two clusters for breastfed and weaned samples. There is a slight separation between the red and blue clusters, indicating that there is a difference in the metabolite concentrations between weaned and breastfed infants. Looking at the variable map, we can identify the metabolite factors which contribute the most to the two largest principal components. Since only the three largest arrows were plotted in the map, we know that these metabolites have the highest contributions to PC1 and PC2. These three arrows represent 4-hydrophenyllactic acid, phenyllactic acid, and indoleacetic acid. This means that these three metabolites can be found in higher fecal concentrations in breastfed infants.

The purpose of figure 1d is to identify which metabolites have the highest correlation to *Bifidobacterium* concentrations using a Spearman's Rank heatmap. After calculating the rho values for correlation, the heatmap visualizes those values based on color. Because lower rho values indicate lower correlation, and higher values represent higher correlation, the lighter the color on the heatmap, the higher the correlation. The figure is separated into three sections based upon the metabolite type. In the tyrosine metabolite heatmap, 4-hydrophenyllactic acid is shown in a light beige color, indicating a high correlation. Of the Phenylalanine metabolites, phenyllactic acid had the highest correlation, and of the Tryptophan metabolites, indoleacetic acid had the highest correlation. This high correlation means that 4-hydrophenyllactic acid, phenyllactic acid, and indoleacetic acid concentrations are correlated with higher *Bifidobacterium* concentrations.

The conclusion of the study found that *Bifidobacterium* convert aromatic amino acids into the lactic acids 4-hydrophenyllactic acid, phenyllactic acid, and indoleacetic acid. This process is made possible by aromatic lactate dehydrogenase (ALDH) contained in the *Bifidobacterium*. There is a positive correlation between aromatic lactic acid concentrations and the concentrations of ALDH containing microbes such as *Bifidobacterium*.

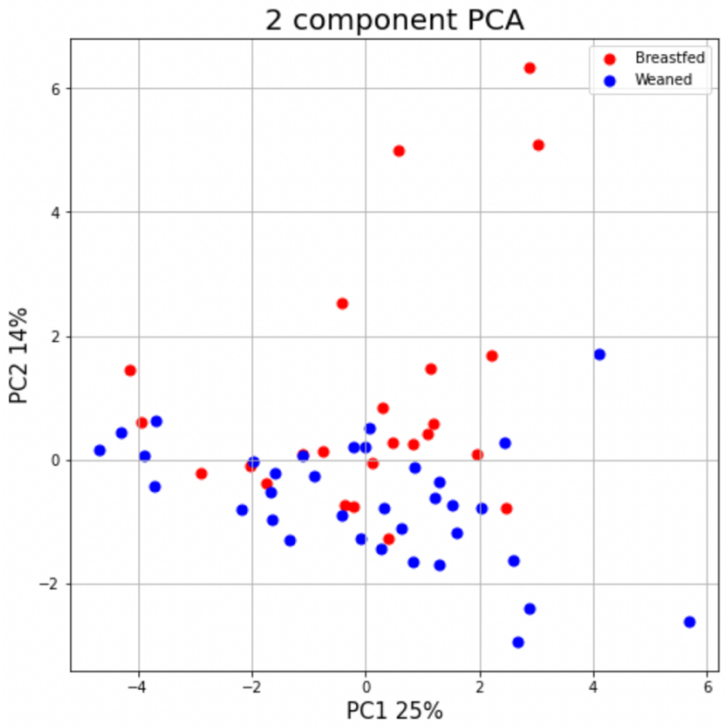
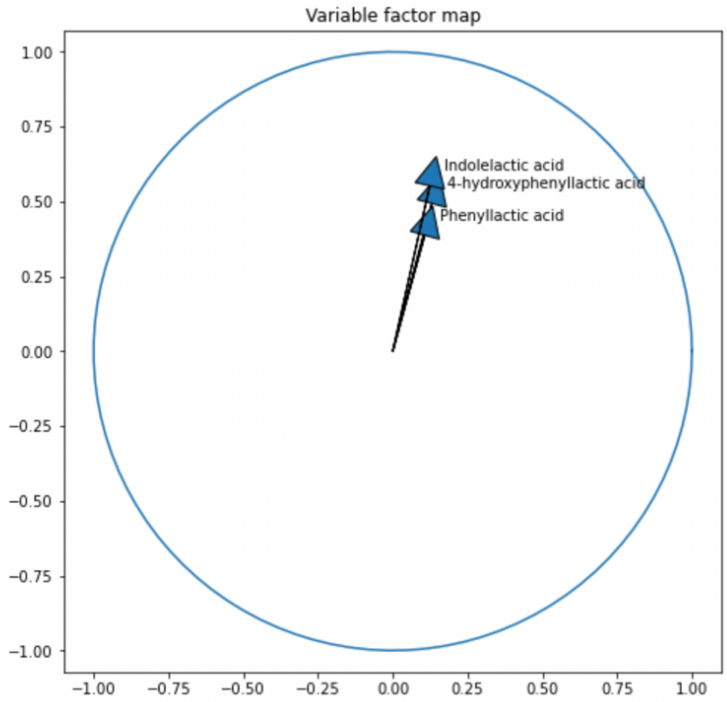


figure 1c



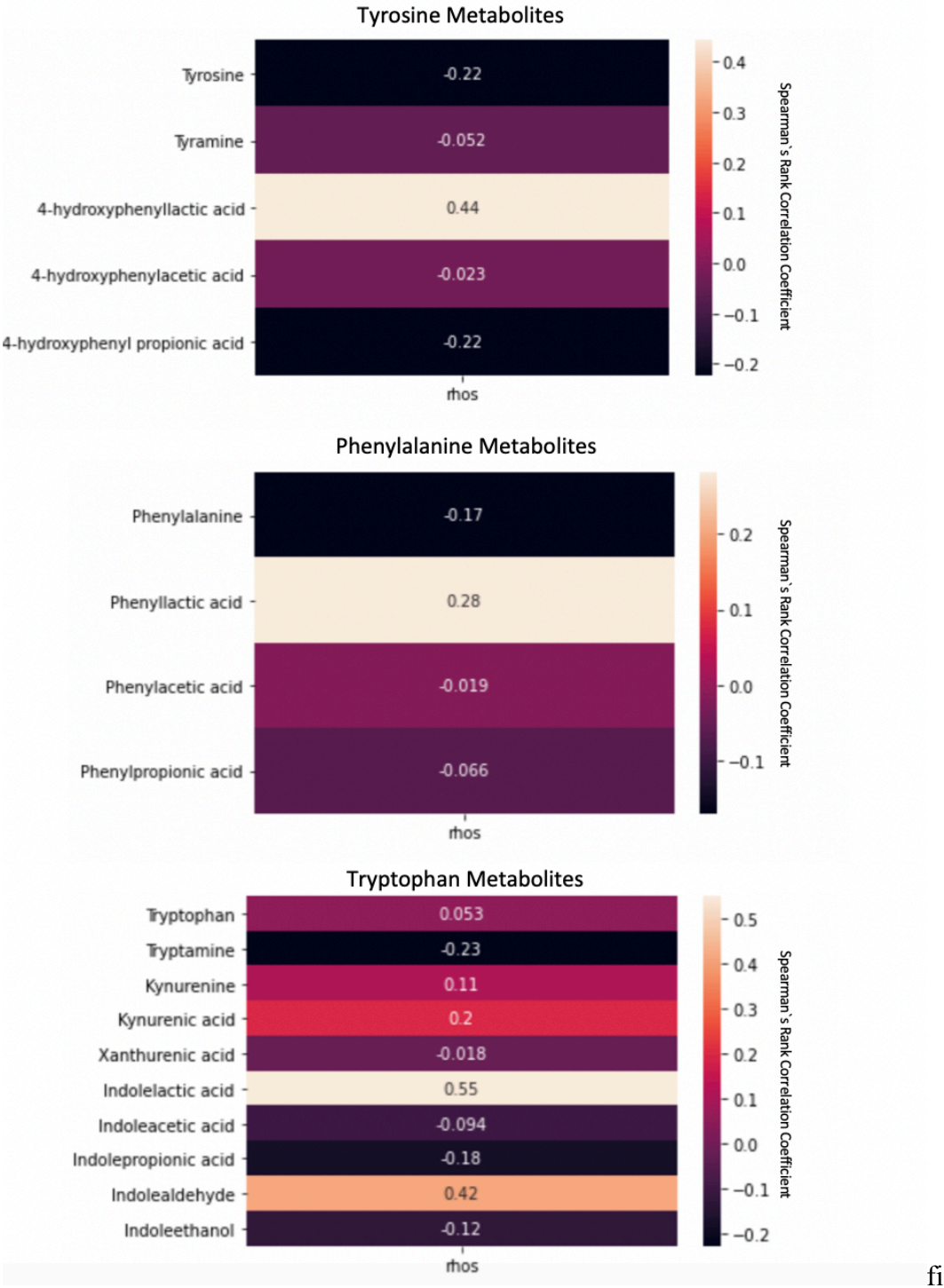


figure 1d

References

1. Laursen, M.F., Sakanaka, M., von Burg, N. *et al. Bifidobacterium* species associated with breastfeeding produce aromatic lactic acids in the infant gut. *Nat Microbiol* **6,** 1367–1382 (2021). https://doi.org/10.1038/s41564-021-00970-4